Overview of METADAC analysis

Nurfatima Jandarova

Table of contents

1 Original data **1**

2 Sample filters **2**
2.1 Voting eligibility . . . 2
2.2 Non-missing covariates . . . 3
2.3 Standard QC . . . 4
2.4 European ancestry . . . 5

3 Phenotypes **6**
3.1 Voting indicator . . . 6
3.1.1 Political alignment . . . 6

4 Descriptives **7**
4.1 Party affiliation . . . 7
4.2 Cognitive test results . . . 8
4.3 Big 5 personality . . . 9

5 Variant filters **10**

6 Imputation **10**

7 Post-imputation QC **11**

8 GWAS **11**
8.1 GRM . . . 12
8.2 Estimate with BOLT-LMM . . . 13
8.3 Format output . . . 13

References **14**

1 Original data

The genotyping data is provided by the UK Household Longitudinal Study (UKHLS). It is a panel survey of the UK population that started in 2009 and follows the sample members each year since. Each wave contains information on about 40,000 individuals. In waves 2 and 3 (years 2010-2011), the survey participants were also asked for their bio samples, which were later genotyped. The genotype information is available for slightly less than 10,000 individuals. For more information, see Benzeval, Aguirre, and Kumari (2023).

We received the data back in 2020, from the METADAC[1], hence the name. We received the data in two batches: April 2020 and June 2023. In April 2020 we received all the genotype data and most of the requested survey variables. In June 2023 we received variables related to political participation

[1]Since then the data management has shifted back to the UKHLS.

and party affiliations. Due to data protection considerations, the survey data released together with the genotype data cannot be linked with the full survey dataset (i.e., one with 40,000 observations). The individual identifiers in the genotype survey dataset is not the same as IDs in the full survey.

April 2020:

```
01_Data/MDAC-2019-0004-03E-ICHINO_20200406//MDAC-2019-0004-03E-ICHINO_sendout.bed
01_Data/MDAC-2019-0004-03E-ICHINO_20200406//MDAC-2019-0004-03E-ICHINO_sendout.bim
01_Data/MDAC-2019-0004-03E-ICHINO_20200406//MDAC-2019-0004-03E-ICHINO_sendout.dta
01_Data/MDAC-2019-0004-03E-ICHINO_20200406//MDAC-2019-0004-03E-ICHINO_sendout.fam
```

June 2023:

```
01_Data/MDAC-2019-0004-03E-ICHINO_20230623//MDAC-2019-0004-03E-ICHINO_202306_sendout.dta
01_Data/MDAC-2019-0004-03E-ICHINO_20230623//MDAC-2019-0004-03E-ICHINO_202306_sendout.txt
```

The genotype data contains information on more than 500,000 variants and 9,921 individuals (see below number of lines in .bim and .fam files).

```
  518542 01_Data/MDAC-2019-0004-03E-ICHINO_20200406/MDAC-2019-0004-03E-ICHINO_sendout.bim
    9921 01_Data/MDAC-2019-0004-03E-ICHINO_20200406/MDAC-2019-0004-03E-ICHINO_sendout.fam
```

2 Sample filters

The analysis plan specified the following sample inclusion criteria:

> *We will use the following individual inclusion criteria:*
>
> a) *They are of European genetic ancestries*
> b) *They are eligible to vote, i.e. have been age-eligible to vote at least once, and are eligible on grounds of residence/citizenship etc.*
> c) *The control variables specified above are non-missing*
> d) *They were genotyped successfully (genotyping call rate >95%, per chromosome missingness rate < 5%)*
> e) *They passed the cohort-specific standard quality controls, e.g. excluding individuals who are ancestry/heterozygosity outliers in the cohort, individuals whose reported sex does not match their sex derived from their genotypes, duplicates, etc..*

2.1 Voting eligibility

`03_Analysis/clean_1a_sample_variables.qmd`

Voting eligibility is based on citizenship and age (at least 18).

Although the dataset has a variable for citizenship status, it is missing for 97% of the sample. Therefore, I do not use citizenship to identify eligibility.

```
# A tibble: 3 x 3
  ever_citizen            n   freq
  <dbl+lbl>           <int>  <dbl>
1  0 [not mentioned]     93  0.938
2  1 [Mentioned]        217  2.19
3 NA                   9610 96.9
```

The genotype survey dataset has variable for age at the time of wave 3 survey (`c_age_dv`). I first extrapolate this information to other waves. For example, subtracting one year in wave 2, or adding one year in wave 4, etc. Thus, the inclusion criteria is that an individual should be age 18 or over at least once in the years when voting question was asked.

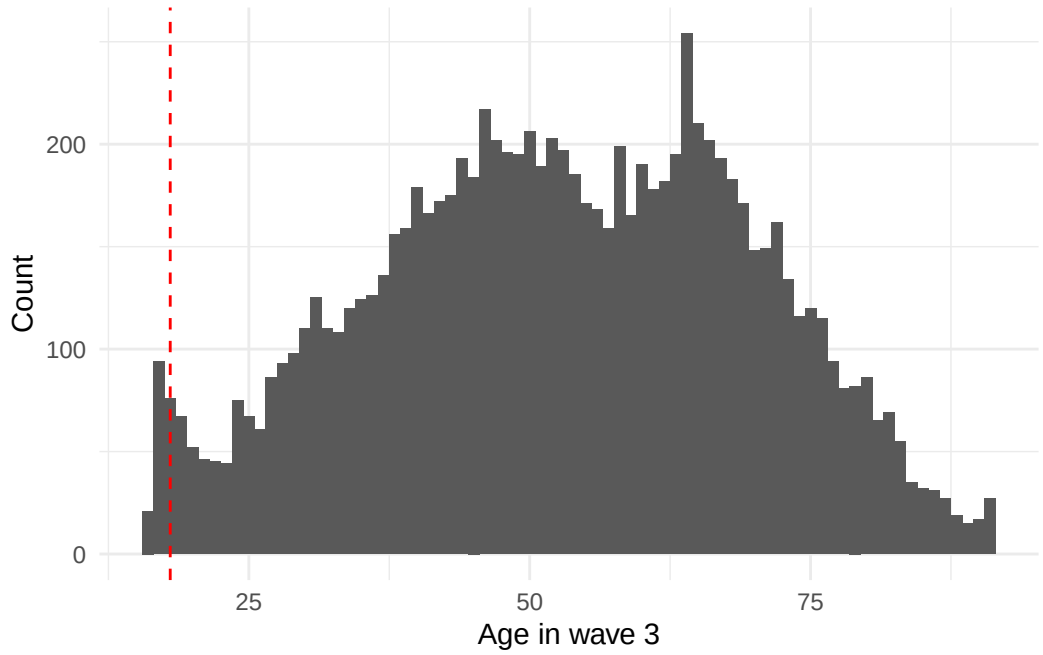


Figure 1: Age distribution in wave 3

Additionally, I use information recorded in the voting indicator itself. For example, here is the tabulation of the voting question in wave 2

```
# A tibble: 4 x 2
  b_vote7                n
  <dbl+lbl>          <int>
1  1 [yes]             3750
2  2 [no]               965
3  3 [can't vote]        94
4 NA                   5111
```

So, individuals ineligible to vote for whatever reason are assigned value 3 in voting indicator. I combine the age limit and voting indicator information in each wave when voting was observed. Then, for each individual, I construct a binary variable whether they were ever eligible to vote (see tabulation below). Some observations in eligibility indicator are missing because they don't have age information in wave 3 and they have never responded to any voting questions. I drop observations who were never eligible to vote and those whose eligibility information is missing.

```
# A tibble: 3 × 2
  ever_eligible     n
          <int> <int>
1             0     5
2             1  9775
3            NA   140
```

2.2 Non-missing covariates

`03_Analysis/clean_1a_sample_variables.qmd`

The covariates used in the analysis are

- genetic principal components
- genotype/imputation batch
- sex
- age

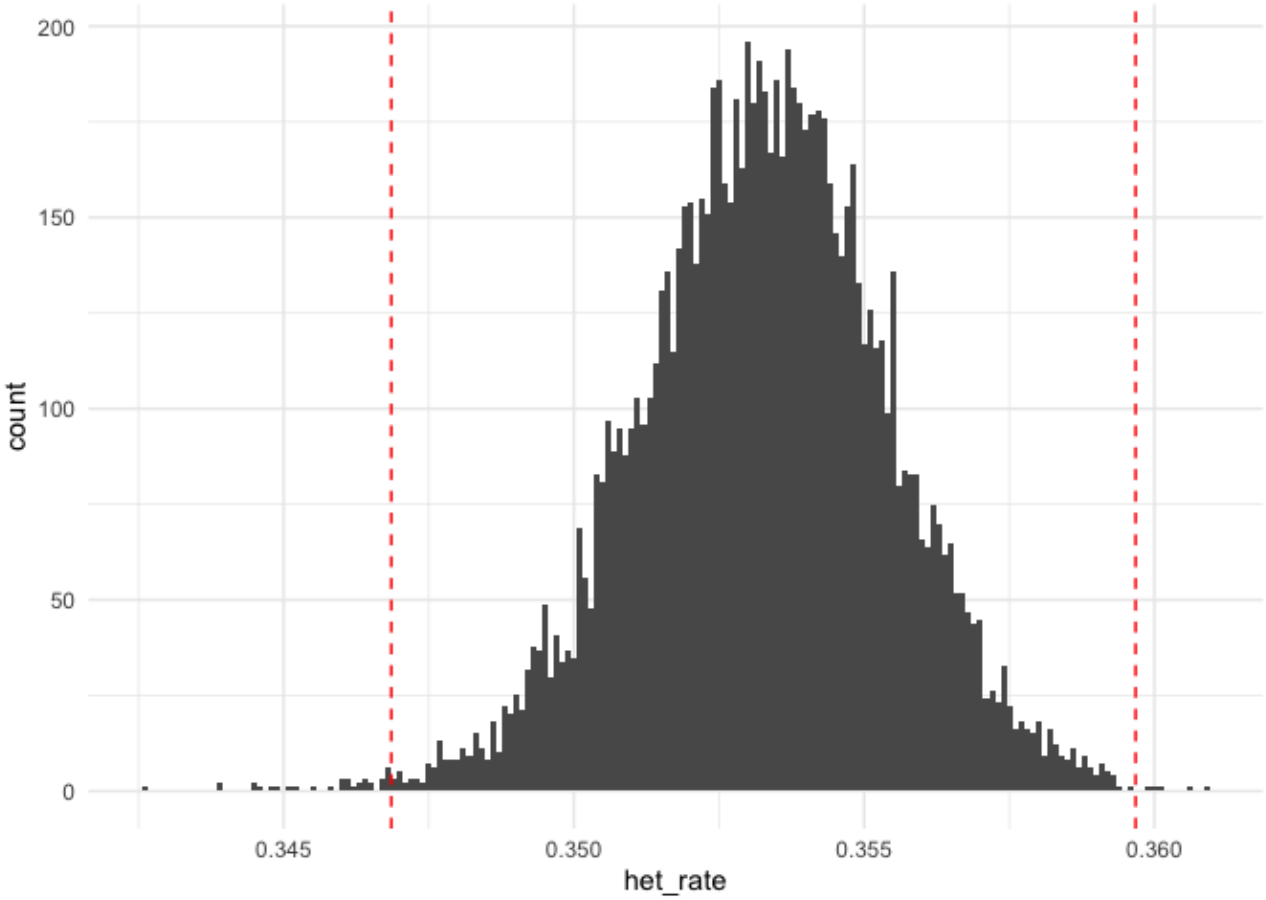
The dataset does not contain any variable for genotype batches. Hence, I remove 218 observations with missing sex and age variables.

2.3 Standard QC

`03_Analysis/clean_1b_sample_genotypes.qmd`

- Remove individuals with more than 5% of genotype information missing
- Remove heterozygosity outliers (outside of 3 sd range around the mean)

```
# A tibble: 2 × 2
  pass      n
  <lgl> <int>
1 FALSE    40
2 TRUE   9517
```



- Remove individuals with mismatch of observed and genetic sex
- Remove related individuals (using `--rel-cutoff 0.025`)

```
# A tibble: 5 × 6
  call  het   sex   rel   pass_all     n
  <lgl> <lgl> <lgl> <lgl>    <int> <int>
1 TRUE  FALSE TRUE  FALSE        0     5
2 TRUE  FALSE TRUE  TRUE         0    35
3 TRUE  TRUE  FALSE TRUE         0     1
4 TRUE  TRUE  TRUE  FALSE        0   785
5 TRUE  TRUE  TRUE  TRUE         1  8731
```

2.4 European ancestry

`03_Analysis/clean_1c_sample_ancestry.qmd`

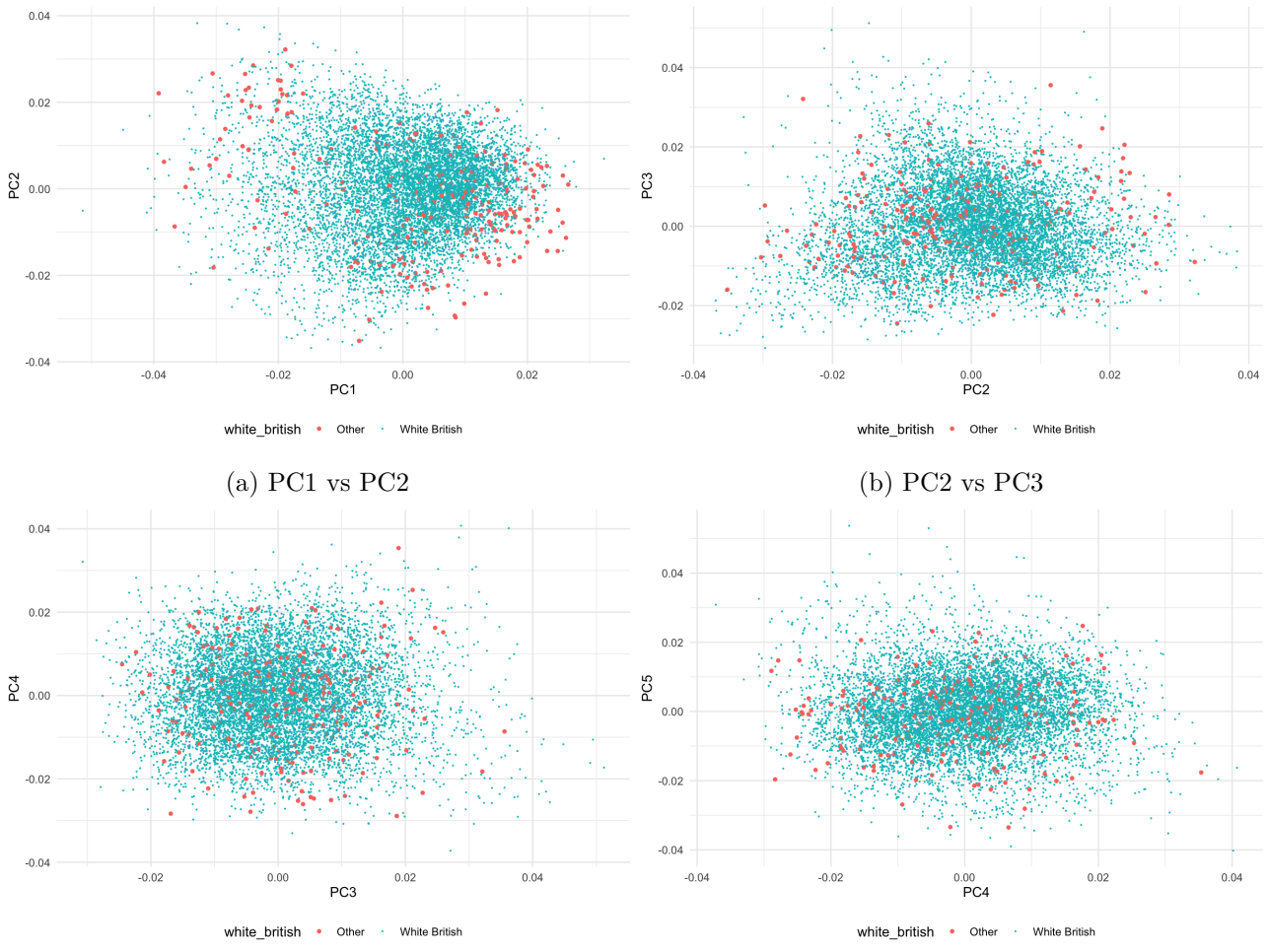
There is no ready-made variable in the dataset that indicates whether someone is from European genetic ancestry. To identify ancestry, I compute genetic PCs and compare them with the ethnicity variables available in the survey data.

I compute the genetic PCs using genotypes of individuals passing the above sample with genotypes standard QC filters (sample call rate $>$ 95%, genotype call rate $>$95%, MAF $\geq$ 1%, HWE p-value $\geq 10^{-6}$ and pruned for LD).

The survey dataset contains information on ethnicities, which I use to construct White British indicator.

Figure 2 plots PC projections by survey ethnicity. So, there don't seem to be clear separation between ethnic groups. I interpret this result as the sample already being predominantly of White European ancestry. This is consistent with the genotyping description in the Understanding Society in Benzeval, Aguirre, and Kumari (2023):

> *At the time, large-scale genotyping was limited to people of White European descent because the reliability of techniques to accurately genotype people varied in different ethnic groups. After data cleaning and other quality control steps, approximately 9,900 samples are available for analysis.*



(a) PC1 vs PC2

(b) PC2 vs PC3

(c) PC3 vs PC4

(d) PC4 vs PC5

Figure 2: PC projections and survey ethnicity

Thus, I keep all individuals surviving the previous filters. Finally, I re-compute genetic principal components in `03_Analysis/clean_1d_genetic_PCs.qmd` and merge first 20 PCs to the working

survey dataset in `01_Data/metadac_gwas.dta`.

Number of individuals passing sample filters

```
8731 01_Data/imputation_files/metadac_clean.fam
```

3 Phenotypes

`03_Analysis/clean_1e_sample_phenotypes.qmd`

I saved phentoype dataset in `01_Data/gwas_files/pheno.txt`, making sure that it follows the format requirements of BOLT-LMM software. Since I residualise the phenotypes with respect to covariates suggested in the analysis plan, I don't need to save the covariates along with phenotypes.

3.1 Voting indicator

The survey dataset contains several indicators about political participation. First, voting behaviour in the most recent general election. This is **first-order election in the UK**. During the sample coverage there were four general elections: May 2010 (`b_vote7`), May 2015 (`g_vote7`), June 2017 (`h_vote7, i_vote7, j_vote7`), December 2019 (`k_vote7, l_vote7`)[2]. I recode these indicators to binary variables where 1 means individual voted in GE, 0 means she did not vote and missing otherwise (i.e., if original variable is missing or if ineligible to vote in that election).

```
pos  variable label
1299 b_vote7  voted in last general election
1361 g_vote7  Voted in last general election
1363 h_vote7  Voted in last general election
1386 i_vote7  Voted in last general election
1396 j_vote7  Voted in last general election
1404 k_vote7  Voted in last general election
1422 l_vote7  Voted in last general election
```

The analysis plan suggests that multiple voting observations per person should be aggregated as follows.

> Since in some cases, there will be data on multiple elections for the same individual, we will aggregate data at the individual level. To accomplish this, and remove as much noise in the measurement as possible, we will use the following procedure to define the phenotype measures. First, for each election, we will linearly regress the binary turnout measure on the covariates listed in Section 5. . Then, we calculate the standardized residuals of this regression. Finally, we take the average of these standardized residuals across elections for each individual. This gives a single measure per individual. Note that if only data on i.e. year of birth is available (rather than month or date), age at time of the election should be set to the "most likely" match by the half-year threshold: *election year minus birth year* if the election was held in the second half of the year, and *election year minus birth year minus 1* if the election was held in the first half of the year.

I follow this algorithm and save the standardised residuals in the `avg_res_voted_high` variable.

3.1.1 Political alignment

In addition to actual voting behaviour, the survey asks all adult participants whether they feel close to or support any political party. This variable is asked almost every wave.

```
pos  variable label
1293 b_vote1  supports a particular political party
```

[2]You can notice that survey methodology changed in 2017. Instead of asking the voting question immediately after the corresponding general election, they started asking the question in each wave. As a result, the answers in recent waves may refer to elections held 2-3 years ago.

```
1314 c_vote1  supports a particular political party
1324 d_vote1  supports a particular political party
1332 e_vote1  supports a particular political party
1347 f_vote1  supports a particular political party
1355 g_vote1  Supports a particular political party
1365 a_vote1  supports a particular political party
1380 i_vote1  Supports a particular political party
1390 j_vote1  Supports a particular political party
1398 k_vote1  Supports a particular political party
1416 l_vote1  Supports a particular political party
```

```
pos  variable label
1294 b_vote2  closer to one political party than others
1315 c_vote2  closer to one political party than others
1325 d_vote2  closer to one political party than others
1333 e_vote2  closer to one political party than others
1348 f_vote2  closer to one political party than others
1356 g_vote2  Closer to one political party than others
1366 a_vote2  closer to one political party than others
1381 i_vote2  Closer to one political party than others
1391 j_vote2  Closer to one political party than others
1399 k_vote2  Closer to one political party than others
1417 l_vote2  Closer to one political party than others
```

I follow similar strategy as with actual voting indicators. I convert these to binary variables in each election, aggregate across waves, and save standardised residuals in `avg_res_aligned` variable.

4 Descriptives

`03_Analysis/clean_1e_sample_descriptives.qmd`

The survey variables are saved in `01_Data/METADAC.descriptives.20241220.xls`. In particular, it contains voting and alignment variables from each wave (and corresponding ages), party choices, cognitive test results and Big5 personality test results.

4.1 Party affiliation

Similar to political participation variables, party choices also are divided into two groups of variables. First, is the party actually voted for in the last general election. I create four binary variables for Conservative, Labour, LibDem and other party choices. The binary indicators are missing if the original party choice variable is missing or if individual was ineligible to vote in that election.

Second, is the party one would vote for tomorrow or party one feels closest to[3]. I combine this information into one variable and call it party alignment. This variable is also asked almost every wave. Again, I construct four binary variables in each wave corresponding to alignment with Conservative, Labour, LibDem or other party.

I use different aggregation algorithm with party information: I create binary variables recording if they have ever voted for/aligned with a given party.

	Mean	SD	N
ever_cons_aligned	0.52	0.50	8714
ever_lab_aligned	0.48	0.50	8714

[3]These variables are mutually exclusive, i.e., if someone indicated party she would vote for tomorrow, then party closest to is missing, and vice versa.

	Mean	SD	N
ever_libdem_aligned	0.25	0.44	8714
ever_other_aligned	0.58	0.49	8714
ever_cons_voted_high	0.54	0.50	6680
ever_lab_voted_high	0.39	0.49	6680
ever_libdem_voted_high	0.22	0.41	6680
ever_other_voted_high	0.19	0.39	6680

4.2 Cognitive test results

The survey administered cognitive tests to adult respondents in wave 3:

- word recall (immediate `c_cgwri_dv` and delayed `c_cgwrd_dv`),
- serial 7 subtraction `c_cgs7cs_dv`,
- numbers series `c_cgns1sc6_dv` and `c_cgns2sc6_dv`,
- verbal fluency `c_cgvfc_dv`, and
- numeric ability `c_cgna_dv`.

The dataset contains variables with counts of correct answers to each of the test. I combine these counts into cognitive ability score using confirmatory factor analysis (CFA). Before doing this, it helps to first estimate the CFA in the full survey, i.e., including non-genotyped individuals. After CFA estimation, I predict the scores in the survey dataset. Before running the CFA I standardise the test results in each birth cohort and gender cell to account for age and gender differences in results.

```
lavaan 0.6-18 ended normally after 31 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        20

                                                  Used       Total
  Number of observations                         27485       30230
  Number of clusters [c_psu]                      5938
  Sampling weights variable                 c_indinub_xw

Model Test User Model:
                                              Standard      Scaled
  Test Statistic                               720.205     508.175
  Degrees of freedom                                 7           7
  P-value (Chi-square)                           0.000       0.000
  Scaling correction factor                                  1.417
    Yuan-Bentler correction (Mplus variant)

Parameter Estimates:

  Standard errors                           Robust.cluster
  Information                                     Observed
  Observed information based on                    Hessian

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Mem =~
    c_cgwri_std       1.000
    c_cgwrd_std       0.956    0.015   64.486    0.000
```

```
  Nmb =~
    c_cgs7cs_std      1.000
    c_cgna_std        1.534    0.036   42.737    0.000
  G =~
    c_cgnssc6_std     1.000
    c_cgvfc_std       0.680    0.016   43.368    0.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Mem ~
    G                 0.651    0.014   45.882    0.000
  Nmb ~
    G                 0.652    0.017   38.938    0.000

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
 .Mem ~~
   .Nmb               0.000

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
   .c_cgwri_std       0.068    0.008    8.810    0.000
   .c_cgwrd_std       0.074    0.009    8.746    0.000
   .c_cgs7cs_std      0.041    0.007    5.603    0.000
   .c_cgna_std        0.076    0.008    9.369    0.000
   .c_cgnssc6_std     0.017    0.008    2.087    0.037
   .c_cgvfc_std       0.062    0.009    6.756    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .c_cgwri_std       0.219    0.009   23.287    0.000
   .c_cgwrd_std       0.309    0.012   25.267    0.000
   .c_cgs7cs_std      0.678    0.014   50.149    0.000
   .c_cgna_std        0.370    0.010   36.562    0.000
   .c_cgnssc6_std     0.549    0.008   65.511    0.000
   .c_cgvfc_std       0.704    0.010   71.990    0.000
   .Mem               0.448    0.011   41.712    0.000
   .Nmb               0.026    0.004    6.216    0.000
    G                 0.425    0.010   41.470    0.000
```

4.3 Big 5 personality

In wave 3, adult participants have also answered Big 5 personality tests. The total scores along the five dimensions are recorded in `c_big5a_dv, c_big5c_dv, c_big5e_dv, c_big5n_dv, c_big5o_dv`. I combine these scores into single big5 personality score using PCA (also run it first in the full survey and predict in the genotyped survey data).

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5
Standard deviation     1.3450 0.9833 0.9227 0.8369 0.8199
Proportion of Variance 0.3618 0.1934 0.1703 0.1401 0.1344
Cumulative Proportion  0.3618 0.5552 0.7255 0.8656 1.0000
```

5 Variant filters

`03_Analysis/clean_2a_variants_genotypes.qmd`

The analysis plan suggests the following

> Genotypes on all autosomal SNPs should be imputed against the 1000 Genomes Phase 3 (1000 Genomes Project Consortium et al. 2015), Haplotype Reference Consortium (McCarthy et al. 2016), or TOPmed reference panels (Taliun et al. 2021). We recommend SNPs with the following properties to be excluded prior to imputation:
>
> - Call rate $< 95\%$
> - Hardy Weinberg Equilibrium test P-value $< 10^{-6}$
> - MAF $< 1\%$
> - SNPs with alleles not matching the reference panel
> - SNPs with allele frequencies significantly deviating from the reference panel
> - SNPs with duplicated base pair position
> - palindromic[4] SNPs with MAF>0.4
>
> These filters may change or other additional filters may be appropriate depending on cohort characteristics. If in doubt, please contact us for recommendations.

I use PLINK2 to apply the call rate, HWE and MAF filters.

I use the HRC 1.1 as a reference panel. The list of the HRC 1.1 variants can be downloded from ftp://ngs.sanger.ac.uk/production/hrc/HRC.r1-1/HRC.r1-1.GRCh37.wgs.mac5.sites.tab.gz.

For the rest of the filters I use HRC or 1000G Pre-imputation Checks software written by Will Rayner. This tool is recommended by the Michigan Imputation Server 2 guidelines. It restricts the data to overlapping variants (between analysis data and reference panel) and enforces allele alignment (swapping or flipping if necessary). It also removes variants with mismatched alleles and palindromic SNPs with MAF > 0.4. Finally, it splits the cleaned genotypes by chromosomes, converts them to .vcf.gz format and sorts the variants by basepair position.

6 Imputation

For imputation, I use Michigan Imputation Server 2 with HRC 1.1 as a reference panel. That is, imputation is run on the specialised server, not on a local computer. The server requires registration, but is free to use. I upload the cleaned .vcf.gz files to the server, specify the reference panel and submit the job. Since I have run the data through the Pre-Imputation check tool in the previous step, the QC on the server passes easily. The imputation itself takes less than 24 hours to be completed. The results can then be downloaded from the website[5]. Since the files are large (a little less than 1TB), I downloaded them to the Minnesota Supercomputer Institute (MSI) working folder (`/home/rustich0/njandaro/Voting_GWAS/01_Data/imputation_files/download`).

Here is the QC report generated by the Michigan Imputation Server.

Parameter	Value
Samples	8731
Chromosomes	1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 3 4 5 6 7 8 9
SNPs	260486
Chunks	153
Datatype	unphased

[4] These are SNPs with alleles AT, TA, GC or CG.

[5] There is a specific time window within which the results can be downloaded. After the window is passed, the files are deleted from the Michigan Imputation Server.

Parameter	Value
Build	hg19
Reference Panel	hrc-r1.1 (hg19)
Population	eur
Phasing	beagle
Mode	imputation
Statistics	
Alternative allele frequency > 0.5 sites	81,197
Reference Overlap	100.00 %
Match	260,486
Allele switch	0
Strand flip	0
Strand flip and allele switch	0
A/T, C/G genotypes	0
Filtered sites	
Filter flag set	0
Invalid alleles	0
Multiallelic sites	0
Duplicated sites	0
NonSNP sites	0
Monomorphic sites	0
Allele mismatch	0
SNPs call rate < 90%	0
Excluded sites in total	0
Remaining sites in total	260,486
Warning	1 Chunk(s) excluded: < 20 SNPs (see chunks-excluded.txt for details).
Remaining chunk(s)	152

Figure 3 shows the densities of frequencies falling into each part. The first 5000 points from areas of lowest regional densities will be plotted.

7 Post-imputation QC

`03_Analysis/clean_2c_variants_imputedQC.qmd`

The post-imputation QC has similar steps as in Variant filters, with an addition of imputation quality filter

- Imputation quality: $R^2 \geq 0.7$
- MAF $\geq 1\%$
- Call rate $\geq 95\%$
- HWE p-value $\geq 10^{-6}$

Finally, I convert to PLINK .bed format and merge all chromosomes into one file. This script can only be run on the MSI since the raw imputed files are stored only there.

8 GWAS

From the analysis plan

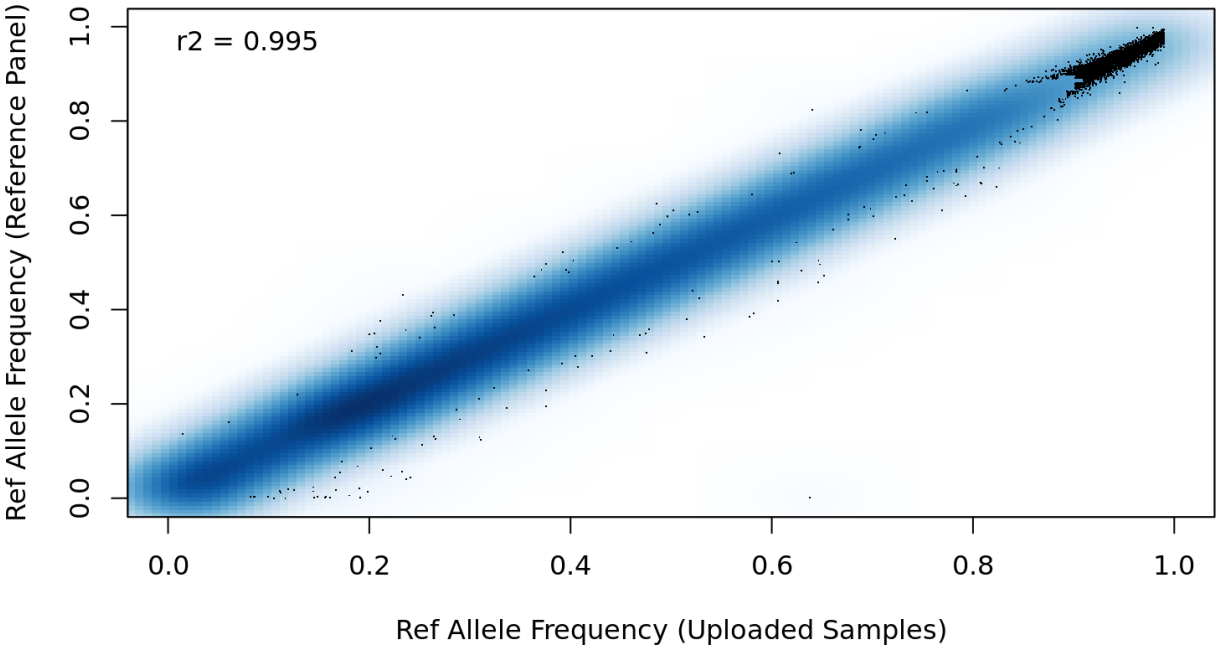


Figure 3: Distribution of allele frequencies in data and reference panel

We recommend conducting mixed linear model based association analysis (MLM) to account effectively for relatedness and population stratification. This method uses a maximum likelihood approach to estimate the following model:

$y = a + bx + g + e$

where y is the phenotype, a is the mean term, b is the additive effect (fixed effect) of the candidate SNP to be tested for association, x is the SNP genotype, g is the polygenic effect (random effect) i.e. the accumulated effect of all SNPs as captured by the genetic relatedness matrix (GRM) and e is the residual.

MLM analysis can be performed in software programs such as GCTA (Yang et al., 2010), BOLT-LMM, FaST-LMM, GEMMA. We recommend calculating the GRM using all genotyped (not imputed) SNPs after restricting individuals to the analysis sample and applying SNP-level quality control filters (e.g. call rate > 95%, MAF>1%, HWE P-value>10^{-6}).

We recommend using the reference allele ***dosage*** *(not the allele count) as the genotype measure in the GWAS. Hard called genotypes do not account for imputation uncertainty.*

8.1 GRM

`03_Analysis/gwas_1a_estimate_grm.qmd`

This script can be run on a local computer. As recommended, I only use genotyped calls passing previous sample and variant filters. These are conveniently saved in

```
01_Data/imputation_files//metadac_clean.bed
01_Data/imputation_files//metadac_clean.bim
01_Data/imputation_files//metadac_clean.fam
01_Data/imputation_files//metadac_clean.frq
```

I also prune them for LD using `--indep-pairwise 500kb 0.2` option of PLINK2. This returns a list

of approximately independent SNP rsid's. In principle, this list of SNPs is sufficient (I don't actually need to compute GRM matrix, but I do).

Number of SNPs for GRM computation

```
74959 01_Data/gwas_files/metadac_grm.prune.in
```

8.2 Estimate with BOLT-LMM

`03_Analysis/gwas_1b_run_bolt.qmd`

As suggested by Tobias, the latest version of BOLT-LMM that runs on MSI is 2.3. I installed it on MSI in `09_Software` folder. This is the code that was executed on MSI.

```
09_Software/BOLT-LMM_v2.3/bolt \
  --lmm \
  --bfile 01_Data/imputation_files/metadac_imputed_clean \
  --phenoFile 01_Data/gwas_files/pheno.txt \
  --phenoCol avg_res_voted_high \
  --LDscoresFile 09_Software/BOLT-LMM_v2.3/tables/LDSCORE.1000G_EUR.tab.gz
  --modelSnps 01_Data/gwas_files/metadac_grm.snps
  --statsFile 01_Data/gwas_files/METADAC.HIGH.stats.20241219
  --verboseStats
```

Number of SNPs in GWAS output

```
 6447248 01_Data/gwas_files/METADAC.HIGH.stats.20241219
```

8.3 Format output

`03_Analysis/gwas_1c_format_output.qmd`

The GWAS summary statistics should be formatted as follows

Variable name (case sensitive!!)	Description
SNPID	SNP identifier (e.g. rs number, CHR:BP, CHR:BP:A1:A2)
CHR	Chromosome
BP_b37	GRCh37 base pair position (if GRCh38 positions are available instead, name the column BP_b38)
EFFECT_ALLELE	Coded allele, also called modeled allele (A/C/G/T/R/I/D) In example of A/G SNP in which AA=0, AG=1 and GG=2, the coded allele is G
OTHER_ALLELE	The other allele
EAF	Effect allele frequency
BETA	Beta estimate from genotype-phenotype association, at least 5 decimal places – 'NA' if not available
SE	Standard error of beta estimate, to at least 5 decimal places – 'NA' if not available
P	*P-value* of test statistic – 'NA' if not available
N	Sample size
INFO	Imputation accuracy for imputed SNPs
HWE_PVAL	HWE P-value for genotyped SNPs
CALLRATE	Genotyping call rate for genotyped SNPs

Most of it is already saved by BOLT-LMM. I add

- sample size (extracting it from BOLT-LMM log file, since final sample size was slightly smaller than 8731 individuals)
- imputation score (extracting R2 from .info.gz files downloaded from the Michigan Imputation Server 2)
- HWE p-values and call-rates (after running corresponding PLINK2 commands on clean genotyped data).

The output is saved in

`01_Data/gwas_files/METADAC.HIGH.association-results.20241219.txt`

References

Benzeval, Michaela, Edith Aguirre, and Meena Kumari. 2023. "Understanding Society: Health, Biomarker and Genetic Data." *Fiscal Studies* 44 (4): 399–415. https://doi.org/10.1111/1475-5890.12354.